Background:

Environmental, Social, and Governance (ESG) reporting has rapidly evolved into a crucial mechanism for demonstrating corporate sustainability, managing non-financial risks, and meeting stakeholder expectations. et, organizations operating across diverse sectors often struggle with fragmented ESG data. Different frameworks and standards—such as IFRS, TCFD, SASB, and GRI—each define their own sets of metrics, rendering comprehensive, comparable ESG disclosures a complex endeavor. The outcome is a patchwork of inconsistent indicators, repetitive reporting practices, and substantial obstacles to meaningful cross-company or cross-industry comparisons.

Moreover, ESG datasets themselves can be vast. A single company might track hundreds of metrics for climate-related disclosures (e.g., carbon emissions, water usage), social indicators (e.g., employee health & safety, diversity), and governance factors (e.g., board independence, executive pay ratios). High-dimensional data often impede meaningful insights; analysts face the challenge of determining which subset of metrics truly captures an organization’s sustainability risk profile or differentiates leaders from laggards in ESG performance. Traditional methods for dimensionality reduction or scoring can exacerbate the confusion if they do not account for the semantic overlaps and distinctions among metrics.

Recent research underscores the need for more robust, standardized approaches to ESG data. Yu et al. (2024) highlight that companies frequently rely on ad hoc spreadsheets and unstructured repositories, which complicates data management and inhibits transparent ESG reporting. Their solution involves **an ontology-driven architecture** to unify definitions of ESG metrics through a knowledge graph, enabling consistent terminology and clearer traceability—attributes critical for compliance and comparability. Nevertheless, while this approach powerfully harmonizes ESG data, it only partially addresses the deeper analytical questions of *which* metrics matter most and *how* to compress large numbers of indicators into interpretable dimensions for decision-making.

On the other hand, the scoring of ESG components itself often lacks consistency. Senadheera et al. (2021), for instance, illustrate how scores assigned to the environmental pillar can be inconsistent or even contradictory among various rating agencies. Some firms with sizable carbon footprints may still receive high “E” scores, and coverage differences between providers blur the line between genuinely superior environmental performance and more superficial disclosure strategies. This fragmented scoring landscape demonstrates that raw ESG data—even if neatly organized—can fail to offer meaningful, apples-to-apples comparisons if not subjected to a transparent and unifying analytical process.

Against this backdrop, our project proposes **an Ontology-Enhanced PCA** approach on an ESG dataset. By integrating the semantic clarity of an ontology—Similar to the framework proposed by Yu et al. (2024)—we ensure that metrics are consistently defined across multiple standards and subdivided according to logical ESG categories. This ontology layer alleviates terminological mismatches (e.g., “Greenhouse Gas Emissions” vs. “CO₂ Equivalent Scope 1”), which is vital for any cross-framework integration. Subsequently, **Principal Component Analysis (PCA)** tackles the high dimensionality by extracting core components that capture the bulk of the variance in ESG risk or opportunity factors. Crucially, the ontology ensures interpretability: when PCA reveals a principal component driven largely by carbon intensity, energy consumption, and waste reduction policies, analysts will be able to map that component back to standard ESG themes (such as “Environmental Risk”).

In this way, the proposed methodology addresses the dual challenges of **semantic fragmentation** and **dimensional complexity** in ESG data. It unifies multiple frameworks into a shared ontology, then leverages PCA to highlight the most significant ESG drivers. By doing so, we build upon both the semantic standardization insights provided by Yu et al. (2024) and the critique of inconsistent scoring systems outlined by Senadheera et al. (2021), ultimately striving for a more robust, transparent, and interpretable ESG analysis pipeline.